



**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2019

On-the-fly calibration in computerized adaptive testing

Verschoor, Angela ; Berger, Stéphanie ; Moser, Urs ; Kleintjes, Frans

Abstract: Research on Computerized Adaptive Testing (CAT) has been rooted in a long tradition. Yet, current operational requirements for CATs make the production a relatively expensive and time consuming process. Item pools need a large number of items, each calibrated with a large degree of accuracy. Using on-the-fly calibration might be the answer to reduce the operational demands for the production of a CAT. As calibration is to take place in real time, a fast and simple calibration method is needed. Three methods will be considered: Elo chess ratings, Joint Maximum Likelihood (JML), and Marginal Maximum Likelihood (MML). MML is the most time consuming method, but the only known method to give unbiased parameter estimates when calibrating CAT data. JML gives biased estimates although it is faster than MML, while the updating of Elo ratings is known to be even less time consuming. Although JML would meet operational requirements for running a CAT regarding computational performance, its bias and its inability to estimate parameters for perfect or zero scores makes it unsuitable as a strategy for on-the-fly calibration. In this chapter, we propose a combination of Elo rating and JML as a strategy that meets the operational requirements for running a CAT. The Elo rating is used in the very beginning of the administration to ensure that the JML estimation procedure is converging for all answer patterns. Modelling the bias in JML with the help of a relatively small, but representative set of calibrated items is proposed to eliminate the bias.

DOI: https://doi.org/10.1007/978-3-030-18480-3_16

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-179441>

Book Section

Published Version



The following work is licensed under a Creative Commons: Attribution 4.0 International (CC BY 4.0) License.

Originally published at:

Verschoor, Angela; Berger, Stéphanie; Moser, Urs; Kleintjes, Frans (2019). On-the-fly calibration in computerized adaptive testing. In: Veldkamp, B P; Sluijter, C. Theoretical and Practical Advances in Computer-based Educational Measurement. Cham: Springer, 307-323.

DOI: https://doi.org/10.1007/978-3-030-18480-3_16

Chapter 16

On-the-Fly Calibration in Computerized Adaptive Testing



Angela Verschoor, Stéphanie Berger, Urs Moser and Frans Kleintjes

Abstract Research on Computerized Adaptive Testing (CAT) has been rooted in a long tradition. Yet, current operational requirements for CATs make the production a relatively expensive and time consuming process. Item pools need a large number of items, each calibrated with a large degree of accuracy. Using on-the-fly calibration might be the answer to reduce the operational demands for the production of a CAT. As calibration is to take place in real time, a fast and simple calibration method is needed. Three methods will be considered: Elo chess ratings, Joint Maximum Likelihood (JML), and Marginal Maximum Likelihood (MML). MML is the most time consuming method, but the only known method to give unbiased parameter estimates when calibrating CAT data. JML gives biased estimates although it is faster than MML, while the updating of Elo ratings is known to be even less time consuming. Although JML would meet operational requirements for running a CAT regarding computational performance, its bias and its inability to estimate parameters for perfect or zero scores makes it unsuitable as a strategy for on-the-fly calibration. In this chapter, we propose a combination of Elo rating and JML as a strategy that meets the operational requirements for running a CAT. The Elo rating is used in the very beginning of the administration to ensure that the JML estimation procedure is converging for all answer patterns. Modelling the bias in JML with the help of a relatively small, but representative set of calibrated items is proposed to eliminate the bias.

A. Verschoor (✉) · F. Kleintjes
Cito, P.O. Box 1034, 6801 MG Arnhem, Netherlands
e-mail: angela.verschoor@cito.nl

F. Kleintjes
e-mail: frans.kleintjes@cito.nl

S. Berger · U. Moser
Institute for Educational Evaluation, Associated Institute of the University of Zurich,
Wilfriedstrasse 15, 8032 Zürich, Switzerland
e-mail: stephanie.berger@ibe.uzh.ch

U. Moser
e-mail: urs.moser@ibe.uzh.ch

16.1 Introduction

In recent years a comprehensive body of research on computerized adaptive testing (CAT) has been performed. Kingsbury and Zara (1989) refined item selection with a method for content control. Sympton and Hetter (1985) dealt with overexposure, while Revuelta and Ponsoda (1998) treated the subject of underexposure. Eggen and Verschoor (2006) proposed a method for difficulty control.

Still, the process to develop an operational CAT is an expensive and time consuming task. Many items need to be developed and pretested with relatively high accuracy. Data collection from a large sample of test takers is a complex task, while it is crucial to be able to calibrate these test items efficiently and economically (Stout et al. 2003). Item calibration is especially challenging when developing a CAT targeted to multiple school grades (Tomasik et al. 2018), or when the target population is small and participation low. Several rounds of pretesting may be necessary in those cases to arrive at accurate estimates of item characteristics. Unfortunately, a threat of pretesting is that motivational effects can cause disturbances in the estimation process (e.g., Mittelhaeuser et al. 2015).

There are enough reasons to look for alternative procedures that can be employed to diminish the burden of pretesting for use in digital testing, especially in CAT. In this chapter, strategies for on-line calibration that can replace pretesting entirely, or at least to a certain extent, are evaluated. Ideas of extending a calibrated item pool are not new. They range from extending an existing item pool with a limited number of new items – replenishment – to periodic calibration of all items and persons in real time: *on-the-fly calibration*. Until recently the latter possibilities were limited by computational potential and infrastructure. But in view of the increased demand for CAT, on-the-fly calibration is drawing attention.

16.1.1 Replenishment Strategies and On-the-Fly Calibration

Replenishment strategies were developed in order to extend existing item pools or replace outdated items. Thus, a precondition for replenishment strategies is that a majority of items have been previously calibrated. Only a relatively small portion of items can be pretested. Test takers will usually take both operational and pretest items, carefully merged, so that they cannot distinguish between the two types. The responses on the pretest items often are not used to estimate the ability of the test takers. This procedure is referred to as seeding. Usually, but not necessarily, responses for the seeding items are collected and only when a certain minimum number of observations has been reached, the item pool will be calibrated off-line. In on-line methods, the seeding items are calibrated regularly in order to optimize the assignment of the seeding items to the test takers, using provisional parameter estimates of those seeding items.

Stocking (1988) was the first to investigate on-line calibration methods, computing a maximum likelihood estimate of a test taker's ability using the item responses and parameter estimates of the operational items. Wainer and Mislevy (1990) described an on-line MML estimation procedure with one EM iteration (OEM) for calibrating seeding items. All these methods can be applied to replenish the item pool, that is, to add a relatively small set of new items to the pool in order to expand the pool or to replace outdated items.

Makransky (2009) took a different approach and started from the assumption that no operational items are available yet. He investigated strategies to transition from a phase in which all items are selected randomly to phases in which items are selected according to CAT item selection algorithms. Fink et al. (2018) took this approach a step further by postponing the calibration until the end of an assessment period and delaying reporting ability estimates, so that the method is suitable for high stakes testing for specialist target populations like university examinations.

Usually, however, an operational CAT project is not started in isolation. Frequently, several items are being reused from other assessments or reporting must take place in relation to an existing scale. In those circumstances, data from these calibrated items and scales may be used as a starting point to collect data for the CAT under consideration. Using this approach it is interesting to investigate how many and which type of calibrated items are required to serve as reference items. Furthermore, we assume a situation in which instant parameter updating and reporting is a crucial element, and thus the time needed for calibration should be kept at a pre-specified minimum.

16.1.2 *On-the-Fly Calibration Methods*

In this chapter we concentrate on the 1PL or Rasch model (Rasch 1960) for which the probability of giving a correct answer by a test taker with ability parameter θ on an item with difficulty parameter β is given by

$$P(X = 1|\theta) = \frac{e^{(\theta-\beta)}}{(1 + e^{(\theta-\beta)})}. \quad (16.1)$$

For item calibration, we consider three methods:

- a rating scheme by Elo (1978) which has been adopted by the chess federations FIDE and USCF in 1960;
- the JML procedure described by Birnbaum (1968);
- the MML method proposed by Bock and Aitkin (1981).

16.1.2.1 Elo Rating

According to the Elo rating system, player A with rating r_A has an expected probability of

$$E_{AB} = \frac{10^{(r_A - r_B)/400}}{1 + 10^{(r_A - r_B)/400}} \quad (16.2)$$

of winning a chess game against player B with rating r_B . After the game has been played, the ratings of the players are updated according to

$$r'_A = r_A + K(S - E_{AB}) \quad (16.3)$$

and

$$r'_B = r_B - K(S - E_{AB}), \quad (16.4)$$

where S is the observed outcome of the game and K a scaling factor. From Equations (16.2)–(16.4) it can be seen that Elo updates can be regarded as a calibration under the Rasch model, albeit not one based on maximization of the likelihood function. Several variations exist, especially since K has been only loosely defined as a function decreasing in the number of observations. Brinkhuis and Maris (2009) have shown, however, that conditions exist under which the parameters assume a stationary distribution.

It is clear that the Elo rating scheme is computationally very simple and fast, and therefore ideally suited for instantaneous on-line calibration methods. However, little is known about the statistical properties of the method, such as the rate of convergence, or even if the method is capable at all of recovering the parameters at an acceptable accuracy. The Elo rating method is widely used in situations in which the parameters may change rapidly during the collection of responses, such as in sports and games. The method has been applied in the Oefenweb (2009), an educational setting in which students exercise frequently within a gaming environment.

16.1.2.2 JML

JML maximizes the likelihood L of a data matrix, where N test takers have each responded to n_j items with item scores x_{ij} and sum score s_j , where P_j is the probability of a correct response and $Q_j = 1 - P_j$ refers to the probability of an incorrect response. The likelihood function can be formulated as

$$L = \prod_{j=1}^N \frac{n_j!}{s_j!(n_j - s_j)!} P_j^{s_j} Q_j^{n_j - s_j}. \quad (16.5)$$

By maximizing the likelihood with respect to a single parameter, while fixing all other parameters, and rotating this scheme over all parameters in turn, the idea is that

the global maximum for the likelihood function will be reached. In order to maximize the likelihood, a Newton-Raphson procedure is followed that takes the form

$$\beta_i^{t+1} = \beta_i^t - \frac{\sum_{j=1}^N P_{ij} - x_{ij}}{\sum_{j=1}^N P_{ij} Q_{ij}} \quad (16.6)$$

for updating item parameter β_i , and

$$\theta_j^{t+1} = \theta_j^t + \frac{\sum_{i=1}^{n_j} P_{ij} - x_{ij}}{\sum_{i=1}^{n_j} P_{ij} Q_{ij}} \quad (16.7)$$

for updating person parameter θ_j .

Like the Elo rating, JML is a simple and fast-to-compute calibration method. However, there are a few downsides. In the JML method, the item parameters are structural parameters. The number of them remains constant when more observations are acquired. The person parameters, on the other hand, are incidental parameters, whose numbers increase with sample size. Neyman and Scott (1948) showed in their paradox that, when structural and incidental parameters are estimated simultaneously, the estimates of the structural parameters need not be consistent when sample size increases. This implies that when the number of observations per item grows to infinity, the item parameter estimates are not guaranteed to converge to their true values. In practical situations, JML might still be useful as effects of non-convergence may become apparent only with extreme numbers of observations per item. The biggest issue with the use of JML, however, is a violation of the assumption regarding the ignorability of missing data. Eggen (2000) has shown that under a regime of item selection based on ability estimates, a systematic error will be built up. He has shown that only for MML, violation of the ignorability assumption has no effect.

16.1.2.3 MML

MML is based on the assumption that the test takers form a random sample from a population whose ability is distributed according to density function $g(\theta|\tau)$ with parameter vector τ . The essence of MML is integration over the ability distribution, while a sample of test takers is being used for estimation of the distribution parameters. The likelihood function

$$L = \prod_{j=1}^N \int P(x_j|\theta, \xi, \tau) g(\theta_j|\tau) d\theta_j \quad (16.8)$$

is maximized using the expectation-maximization algorithm (EM). ξ is the vector of item parameters here. EM is an iterative algorithm for maximizing a likelihood function for models with unobserved random variables. Each iteration consists of an

E-step in which the expectation of the unobserved data for the entire population is calculated, and an M-step in which the parameters are estimated that maximize the likelihood for this expectation. The main advantage of the MML calibration method is that violation of the ignorability assumption does not incur biased estimators. A disadvantage, however, is that MML is usually more time consuming than the Elo rating or JML, and thus it might be too slow to be employed in situations where an instantaneous update is needed.

In order to replenish the item pool, Ban et al. (2001) proposed the MEM algorithm, where in the first iteration of the EM algorithm, the parameters of the ability distribution for the population are estimated through the operational items only. However, if the number of operational items is very small, this procedure results in highly inaccurate estimates. For the problem at hand, where not many operational items are available, the MEM method is not applicable, and the regular MML might be considered.

16.1.3 *The Use of Reference Items in Modelling Bias*

Although bias cannot be avoided when calibrating CAT data with JML, previously collected responses for a part of the items can be used to model this bias. We assume that the bias is linear, i.e. that by applying a linear transformation $\hat{\beta}'_i = a\hat{\beta}_i + b$, estimators $\hat{\beta}'_i$ eliminate the bias. When JML is used to calibrate all items in the pool, the reference items, whose previously estimated parameters we trust, can be used to estimate transformation coefficients a and b . Let μ_r and σ_r be the mean and standard deviation of the previously estimated parameters of the reference items, while μ_c and σ_c are the mean and standard deviations of the parameter estimates for the same items, but now taken from the current JML-calibration, then the transformation coefficients are $a = \frac{\sigma_c}{\sigma_r}$ and $b = \frac{\mu_c - \mu_r}{\sigma_r}$. All reference items retain their trusted parameter values, while all new items in the calibration are updated to their transformed estimates.

Even though MML calibrations do not incur bias when applied to CAT data, it is technically possible to follow a similar procedure, but this should not result in substantially improved parameter estimates.

16.1.4 *The Need for Underexposure Control*

A solution must be found for a complication that will generally only occur in the very first phases when few observations are available. For some response patterns, the associated estimates of the JML and MML methods assume extreme values. In the case of perfect and zero scores, the estimations will even be plus or minus infinity. On the other hand, items are usually selected according to the maximization of Fisher information. This means that once an item has assumed an extreme parameter value,

it will only be selected for test takers whose ability estimation is likewise extreme. And thus, those items tend to be selected very rarely. If the parameter estimation is based on many observations, this is fully justified. But if this situation arises when only a few observations have been collected, these items will effectively be excluded from the item pool without due cause.

In previous studies, this situation was prevented by defining various phases in an operational CAT project, starting with either linear test forms (Fink et al. 2018) or random item selection from the pool (Makransky 2009). When the number of observations allowed for a more adaptive approach, a manual decision regarding transition to the next phase was taken. Since we assume a situation in which decisions are to be taken on the fly and have to be implemented instantaneously, manual decisions are to be avoided. Therefore, a rigid system to prevent underexposure must be present so that items with extreme estimations but with very few observations will remain to be selected frequently. Similar to the findings of Veldkamp et al. (2010), we propose an underexposure control scheme based on eligibility function

$$f(n_i) = \begin{cases} X - n_i(X - 1)/M, & n_i < M \\ 1, & \text{else} \end{cases} \quad (16.9)$$

where n_i is the number of observations for item i , M is the maximum number of observations for which the underexposure control should be active, and X is the advantage that an item without any observations gets over an item with more than M observations. This means effectively that there is no transition from different phases for the entire item pool, but each transition will take place on the level of individual items.

Overexposure may further improve the estimation procedure, but since items with only few observations form a larger impediment to the cooperation between estimation procedure and item selection than a loss of efficiency caused by overexposure, overexposure control is out of scope for this study.

16.1.5 A Combination of Calibration Methods

Although Joint Maximum Likelihood (JML) calibration would meet operational requirements for running a CAT in terms of computational performance, its bias and its inability to estimate parameters for perfect or zero scores makes it unsuitable as a strategy for on-the-fly calibration. To overcome this issue, we propose a combination of Elo rating and JML as a strategy that meets the operational requirements for running a CAT. The Elo rating is used in the very beginning of the administration to ensure that the JML estimation procedure is converging for all answer patterns. Modelling the bias in JML with the help of a relatively small, but representative set of calibrated items, is proposed to eliminate the bias in JML. Although Marginal Maximum Likelihood (MML) estimation is not known to give biased estimates, it

may turn out that this method also benefits from modelling the bias and starting with an Elo rating scheme.

Next to the use of well-calibrated reference items there is the need for underexposure control to ensure that we will collect an acceptable minimum of observations for all items to make calibration feasible.

16.2 Research Questions

Research on on-the-fly item calibration in an operational setting for use in a CAT has different aspects. The operational setting requires the process to be accurate enough and quick enough. The main research question is: Does the proposed approach of a combination of Elo and JML yield sufficient accuracy for both item and persons parameters, under operational conditions? If the research question can be positively answered, there is no need for a large calibrated item pool for a CAT implementation. New items will be embedded in a set of reference items, while instantaneous on-the-fly calibration enables us to base test takers' ability estimates on both new and reference items. At the same time, the use of reference items help to improve item parameter estimations.

Elaborating on the main research question, we formulate the following research sub questions:

- Can we model bias in JML by means of reference items? How do we select the reference items in modelling the bias in JML in relation to their scale? How many reference items are required?
- How well does the proposed calibration method recover item and person parameters? Can we give guidelines when to switch from Elo to JML?
- How well does the proposed method perform in terms of computational time?

Two simulation studies were conducted. The first study concentrated on the investigation of the use of reference items in the elimination of bias. The second study concentrated on a comparison of the methods with respect to their ability to recover the item and person parameters, in relation to the computation time needed.

16.3 Simulation Studies

All simulations were based on the same item pool consisting of 300 items, 100 of which had "true" parameters drawn randomly from a standard normal distribution, supplemented by 200 items drawn from a uniform distribution from the interval $(-3.5, \dots, 3.5)$. Furthermore, all simulated students were drawn from a standard normal distribution and took a test with fixed length of 20 items. Simulated students were assigned to the CAT one after another. The first item for each simulated student was drawn randomly from the entire item pool. For selecting the remaining 19 items,

the weighted maximum likelihood estimator (WMLE) by Warm (1989) was used for the provisional ability, while items were selected that had the highest eligibility function value as given in (16.9), multiplied by the Fisher information at the provisional ability. At the start of each run, all item parameter estimates were set to zero, apart from the reference items that retained their trusted value. After each simulated student had finished the test, the item pool was calibrated. Two criteria were used to evaluate the measurement precision of the calibration methods: average bias and root mean square error (RMSE) of the item and person parameter estimators.

The Elo rating was used in the very beginning of the simulation runs. Using the Elo rating, in combination with the underexposure control, effectively prevented the situation that extreme estimations occurred and that items were thus no longer selected. All items that had a minimum of 8 observations and had a response vector consisting of both correct and incorrect answers were calibrated with JML or with MML. At this moment, extreme parameter estimates purely due to chance begin to become somewhat unlikely. For the items not having this threshold yet, the Elo rating was retained.

16.3.1 Use of Reference Items in Elimination of Bias

To validate the reference items approach with JML, we compared the results from JML with and without reference items. After sorting the item pool in ascending difficulty, the reference items were chosen to be evenly spread in an appropriate difficulty range. It may be assumed that such an evenly spread reference set will yield more accurate estimations for the transformation coefficients than, for example, a normal distribution. Items with more extreme difficulties, however, may be selected less frequently and less optimally than items with moderate difficulty. This difference in selection may well influence the extent to which the reference items model the bias. Therefore, two different sets of reference items were simulated in order to investigate which set is able to eliminate the bias best. The first set comprised 20 items with parameters from the interval $(-2.5, \dots, 2.5)$, while in the second set, 20 items from the full range were selected. The item parameters in both reference sets were evenly spread. In summary, we investigated three different conditions in the first simulation study: Two different ranges of reference items (limited vs. full range), while the third condition was the application of JML without reference items. Each condition was repeated 1000 times and each of these runs consisted of 3000 simulated students. This coincided with the moment that the average number of observations is 200 per item, which is a reasonable choice for calibrations under the 1PL model. Furthermore, bias for each item, averaged over the simulation runs, was investigated when 600 simulated students had finished their tests. This moment coincides with the moment that the average number of observations is 40. This point can be seen as being in a transition from the situation in which no information at all is available to the situation in which calibrations under the 1PL model are starting to become stable.

16.3.1.1 Results

Figure 16.1 shows two scatter plots of $\hat{\beta}$ against β at the end of a single run, i.e. when each item had on average 200 observations. The left scatter plot depicts the estimates for the JML, while the right one shows the estimates in case of JML with the limited set of reference items. It shows a typical situation that can be observed in all simulation runs. It can be seen that a linear transformation based on a set of reference items is effective in reducing the bias, even if it can be surmised that the reduction is not perfect for items with extreme parameter values. Nevertheless, the JML without reference items shows bias to such a large extent that it has not been considered any further in the remainder of the simulations.

In Fig. 16.2, the bias in $\hat{\beta}$ is given. The bias for the limited set of reference items is presented in the left graphs, in the right graphs the bias for the full set. The top row shows the bias for each item, averaged over the simulation runs, when 600 simulated students have finished their test, the bottom row at the end of the runs. When an average of 40 observations per item have been collected, the exclusion of extreme items in the reference set lead to a pronounced overcompensation, while using items from the full difficulty range resulted in overcompensation only at the extremes. The situation at an average of 200 observations per item is different: using reference items from the full difficulty range lead to a clear outward bias for the medium items and an inward bias for the extreme items. In the early phases of the simulations, a linear transformation based on the full set gave the best results, but overcompensation caused a situation in which the error in parameters gradually increased. As it may be surmised that items with medium difficulty will be selected more frequently than items with extreme difficulty, it can be expected that the ability estimates will be more accurate when extreme items will be excluded from being reference items. A second

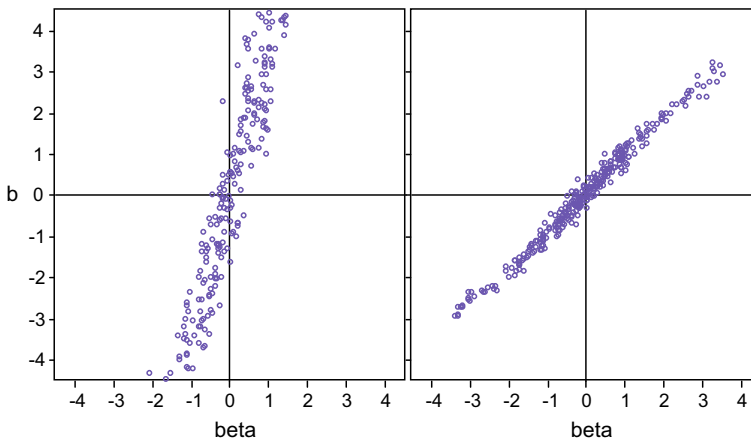


Fig. 16.1 Scatter plot of $\hat{\beta}$ against β for JML without (left) and JML with reference items (right – limited set) at an average of 200 observations per item

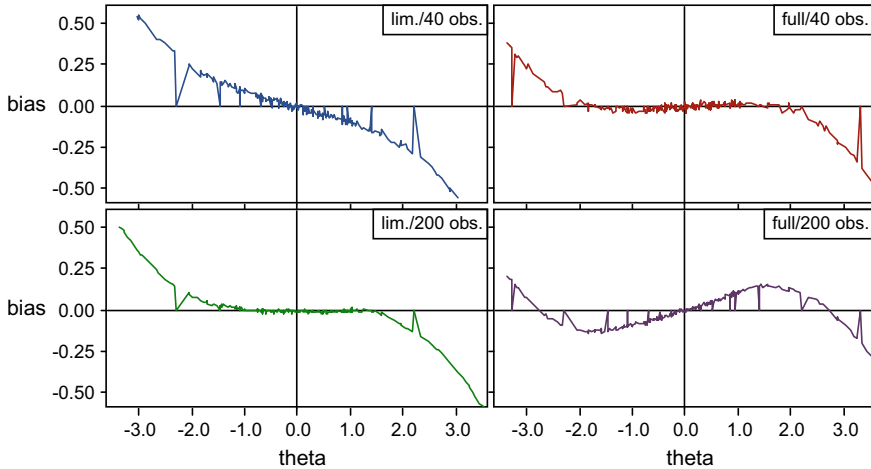


Fig. 16.2 Bias of $\hat{\beta}$ for JML with two different sets of reference items at an average of 40 and 200 observations per item

disadvantage for using extreme items in the reference set is the overcompensation as it can be seen when a considerable number of observations has been collected. Therefore, only JML with a set of reference items selected from a limited difficulty range was considered in the second study. This condition is from here on referred to plainly as JML.

16.3.2 Comparison of the Methods

The settings of the second set of simulations were similar to those of the first set. Like the first set of simulations, we evaluated the bias for person and item parameters after 3000 simulated student, thus making a direct comparison possible. As an extension, we reported the RMSE for the person parameter after each simulated student, as well as the average of the RMSEs for all item parameters. In order to investigate convergence of the methods beyond an average of 200 observations per item, each run consisted of 15,000 simulated students instead of 3000, thus extending the simulation to an average of 1000 observations per item. The conditions that were investigated were:

- Elo rating
- JML
- MML
- MML with reference items.

16.3.2.1 Results

Figure 16.3 shows the bias of $\hat{\beta}$ for the four different simulation conditions, taken after 3000 simulated students. The top left graph shows the results for the Elo method, the top right graph for JML, and bottom left for MML, while in the bottom right graph the results for MML with reference items are presented.

As can be inferred from Eqs. (16.3) and (16.4), the Elo rating update scheme is an effective method in preventing extreme estimations in cases of perfect and zero scores, but in the long run this appears to become a disadvantage. The Elo rating shows a large inward bias in Fig. 16.3, while MML shows an outward bias. The two conditions using reference items showed the best results. JML yielded an increasing inward bias for extreme item parameters, and the use of reference items in the case of MML compensated the bias almost completely.

Recovery of person parameters is equally important as the recovery of item parameters, if not more important. The purpose of a test usually is an evaluation of test takers, item calibration is merely a means to achieve this. Figure 16.4 shows the bias in $\hat{\theta}$. The bias in $\hat{\theta}$ reflected the situation for $\hat{\beta}$: The Elo rating showed substantial inward bias, MML an outward bias, while the use of reference items reduced bias also in the person parameter estimates. In the case of JML, the extreme ability ranges showed some inward bias while a small overcompensation took place for MML.

Figure 16.5 illustrates the development of $RMSE(\hat{\beta})$ during the simulation runs, now averaged over all items. As can be seen, the average $RMSE(\hat{\beta})$ rapidly diminished in the first phases of the process. The Elo rating, combined with the under-exposure control, effectively prevented estimations with extreme values when the first simulated students took their tests. The disadvantage of the Elo rating becomes

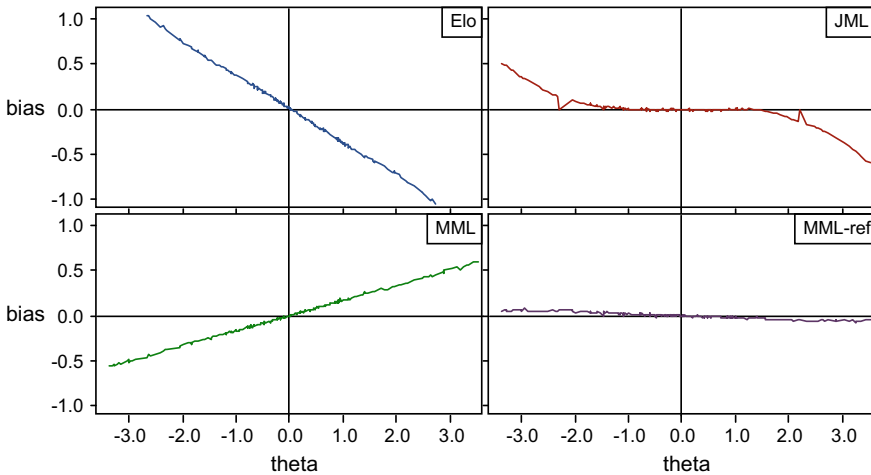


Fig. 16.3 Bias of $\hat{\beta}$ for Elo, JML, MML and MML with reference items at an average of 200 observations per item

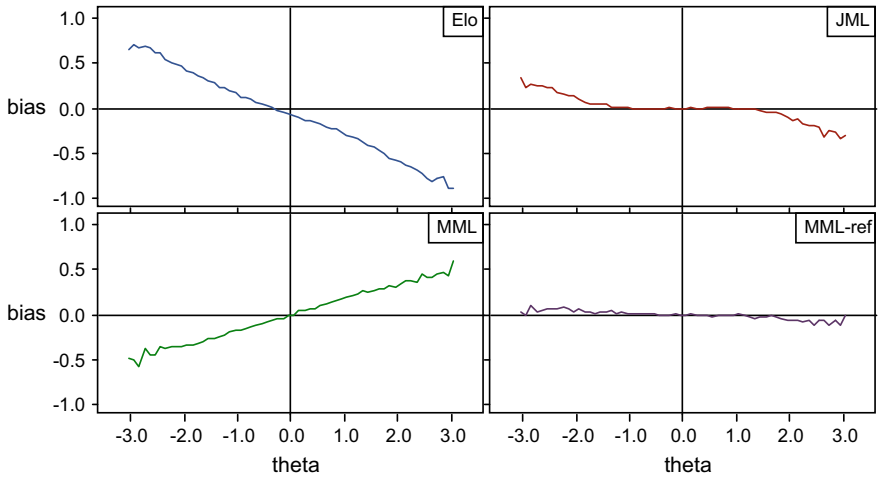


Fig. 16.4 Bias of $\hat{\theta}$ for Elo, JML, MML and MML with reference items at an average of 200 observations per item

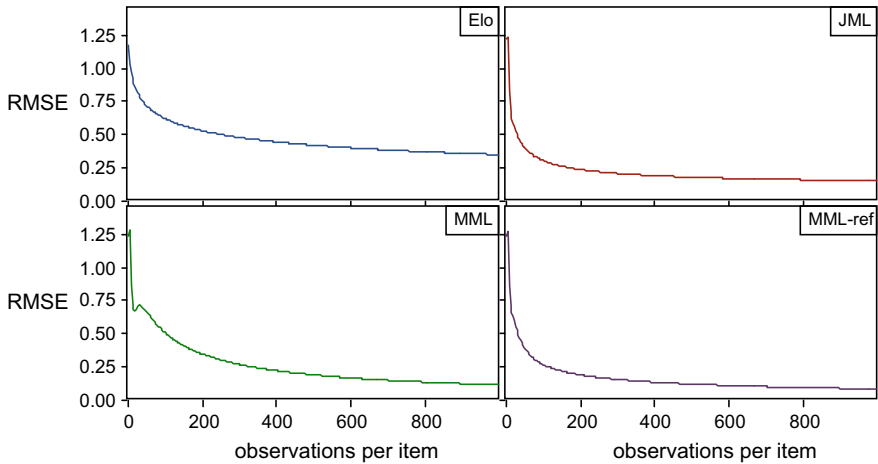


Fig. 16.5 Average $\text{RMSE}(\hat{\beta})$ at increasing numbers of observations

apparent later on in the runs: The Elo rating is slowest to converge of the methods under consideration. At an average of 1000 observations per item, $\text{RMSE}(\hat{\beta})$ was on average 0.351. This situation was reached at an average of approximately 70 observations per item for JML and 60 for MML with reference items. JML converged faster in the initial phase, but with increasing numbers of observations, MML started to outperform JML. MML with reference items had a clear advantage over JML from 50 observations per item. The average $\text{RMSE}(\hat{\beta})$ was 0.383 at that moment.

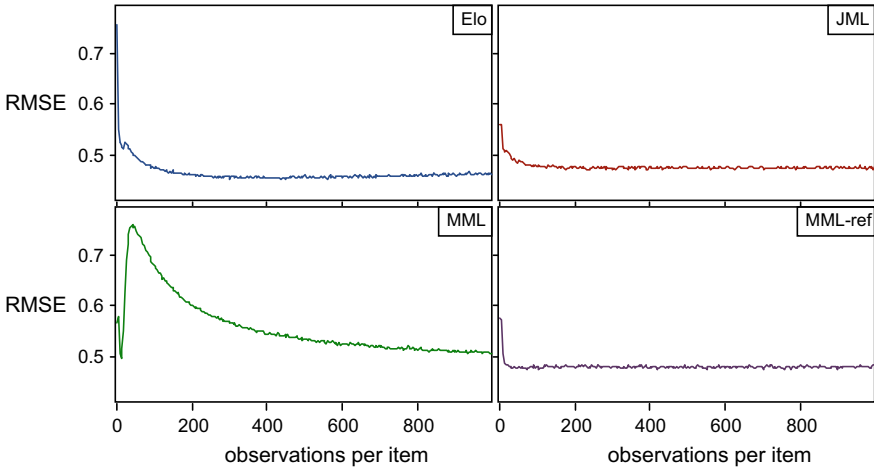


Fig. 16.6 Average $\text{RMSE}(\hat{\theta})$ at increasing numbers of observations

Without using reference items, MML outperformed JML from approximately 500 observations per item, when the average $\text{RMSE}(\hat{\beta})$ was 0.191.

Figure 16.6 presents a different situation. It shows the development of average $\text{RMSE}(\hat{\theta})$. In all conditions, apart from the case of MML without reference items, average $\text{RMSE}(\hat{\theta})$ decreased rapidly. At approximately 50 observations per item, average $\text{RMSE}(\hat{\theta})$ decreased below 0.50 for the Elo ratings. This value was reached at approximately 30 observations per item in the JML and MML with reference items. In the MML condition without reference items, the lowest value for average $\text{RMSE}(\hat{\theta})$ was 0.51 when an average of 1000 observations per item were collected. As a baseline, a simulation was conducted where all item parameters assumed their true value in the item selection algorithm and thus no calibration took place. In this situation, average $\text{RMSE}(\hat{\theta})$ reached a value of 0.48. This value was reached only in the Elo rating condition, albeit because $\hat{\theta}$ showed a large inward bias.

16.3.2.2 Computational Performance

One important issue for on-the-fly calibration in operational settings still remains open: computational performance. Ideally, a calibration method gives highly accurate estimations almost instantly, but in case that such an ideal is not reached, a compromise between accuracy and speed must be made. Unfortunately, in operational settings speed depends not only on the method to be used, but also on other factors like system architecture and interfacing between various software modules. Therefore a complete evaluation of the compromise between accuracy and speed falls outside the scope of this chapter. Nevertheless, measurements of computer times dur-

ing the simulations do give an indication of the relative performance of the calibration methods under consideration.

The Elo rating method is clearly the fastest method, since it makes use of only the responses of one test taker in each cycle and does not consider the entire data matrix. Running on a laptop equipped with an i5-5Y57 CPU at 1.6GHz, one Elo update cycle typically used less than 0.1 ms, regardless of the number of previous responses on the respective items. On the other hand, JML and MML were substantially more computation intensive and showed different time complexity. After 300 simulated students, that is, when on average 20 observations were collected, the average time for one JML calibration cycle took approximately 800 ms. The MML calibration took on average approximately 2.3 s to complete. At 600 simulees, the computation times had increased to 1.6 and 3.8 s for JML and MML, respectively. The computation time increased roughly linearly to the end of the simulation runs, when each JML cycle needed an average of 42 s and a single MML cycle used approximately 140 s to complete. During all phases of the runs, the MML showed negligible differences in computation time between the condition with and without reference items.

16.4 Discussion

From a measurement perspective, MML clearly outperforms JML and Elo ratings, while the use of reference items has an advantage in stabilizing the scale in the early phases of data collection for an item pool while in operation. In the case of JML, the use of reference items should always be considered. But also in the case of MML, reference items are an effective means to eliminate bias. Computational performance, however, heavily depends on external factors such as system architecture and cannot be evaluated in a simulation study as we have conducted. What is clear, though, is that the Elo ratings are extremely fast to compute so that they can be employed in synchronous processes immediately after receiving a response. This is a reason why many large-scale gaming and sports environments use some variant of this method. Furthermore, it is a very useful method in the beginning of data collection processes, when the number of observations is still very low. On the other hand, JML and MML are considerably slower, so that it can be expected that employment of those methods will be deemed feasible only in asynchronous processes. As a consequence, the update of parameters will not be instantaneous but will take a few seconds or more to be effective. In a setting with many thousands of test takers taking the test concurrently, this would be an unwanted situation and the Elo rating system would probably be the method of choice.

When more accurate estimations than those provided by the Elo rating are needed, MML can be used. If reference items are available, JML is an option as well, whereby it could be expected that MML with reference items produce more accurate estimations at the cost of a somewhat heavier computational burden as compared to JML. In both cases it can be expected that item parameter estimations will have stabilized when an average of approximately 200 observations per item have been collected.

At this stage, each consecutive calibration will not change parameter values substantially. Thus, it might be wise not to calibrate the items after each test taker, but to use a less intensive routine. Such a regime could well be an off-line calibration, whereby, for example, item fit statistics can be inspected and actions undertaken when items appear to be defective.

Two important questions were left unanswered in this study: “How many reference items are needed?” and “How many new items can be added?”. It is obvious that the answer to both questions depend on factors such as model fit and accuracy of the reference item estimates. As all responses were simulated under ideal circumstances, that is, all items showed a perfect model fit and all reference items were modelled as having no error, we assume that using two reference items would have been enough to accurately estimate the transformation coefficients, resulting in findings similar to the ones reported here.

The first simulation study, validating the reference items approach, showed that a linear transformation does not completely remove the bias incurred by violation of the ignorability assumptions. A more advanced model of the bias involving a non-linear transformation would probably improve the parameter recovery for the JML calibration, potentially up to the level currently seen with MML using reference items.

Based on findings from the simulation studies, we suggest to apply the following strategy in an on-the-fly calibration in CAT: The entire process of data collection and on-the-fly calibration in an operational CAT setting should start with an Elo rating update scheme in order to avoid extreme parameter estimates, combined with a rather rigorous underexposure control scheme. After this initial phase, a switch to JML using reference items should be made. When increased accuracy is worth the additional computational burden of using MML, a new switch could be made and the on-the-fly calibration can either be slowed down or switched off entirely when parameters appear to be stabilized.

References

- Ban, J., Hanson, B., Wang, T., Yi, Q., & Harris, D. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, 38, 191–212. <https://doi.org/10.1111/j.1745-3984.2001.tb01123.x>.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental scores*. Reading, MA: Addison-Wesley.
- Bock, R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, 46, 443–459. <https://doi.org/10.1007/BF02293801>.
- Brinkhuis, M. & Maris, G. (2009). *Dynamic parameter estimation in student monitoring systems*. Technical Report MRD 2009-1, Arnhem: Cito.
- Eggen, T. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters. *Psychometrika*, 65, 337–362. <https://doi.org/10.1007/BF02296150>.
- Eggen, T., & Verschoor, A. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, 30, 379–393. <https://doi.org/10.1177/0146621606288890>.

- Elo, A. (1978). *The rating of chess Players, past and present*. London: B.T. Batsford Ltd.
- Fink, A., Born, S., Spoden, C., & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, 60, 327–346.
- Kingsbury, G., & Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2, 359–375. https://doi.org/10.1207/s15324818ame0204_6.
- Makransky, G. (2009). An automatic online calibration design in adaptive testing. In D. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*: Reston, VA: Graduate Management Admission Council.
- Mittelhaeuser, M., Béguin, A., & Sijtsma, K. (2015). The effect of differential motivation on irt linking. *Journal of Educational Measurement*, 52, 339–358. <https://doi.org/10.1111/jedm.12080>.
- Neyman, J., & Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16, 1–32. <https://doi.org/10.2307/1914288>.
- Oefenweb B.V. (2009). Math Garden [Computer Software].
- Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Danmarks Pædagogiske Institut.
- Revue, J., & Ponsoda, V. (1998). A comparison of item exposure methods in computerized adaptive testing. *Journal of Educational Measurement*, 35, 311–327. <https://doi.org/10.1111/j.1745-3984.1998.tb00541.x>.
- Stocking, M. (1988). *Scale drift in on-line calibration*. Research Report (pp. 88–28). Princeton: Educational Testing Service. <https://doi.org/10.1002/j.2330-8516.1988.tb00284.x>.
- Stout, W., Ackerman, T., Bolt, D., & Froelich, A. (2003). *On the use of collateral item response information to improve pretest item calibration*. Technical Report (pp. 98–13). Newtown, PA: Law School Admission Council.
- Simpson, J., & Hetter, R. (1985). *Controlling item exposure rates in computerized adaptive testing*. San Diego: Paper presented at the annual conference of the Military Testing Association.
- Tomasik, M., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology*, 9, 1–17. <https://doi.org/10.3389/fpsyg.2018.02245>.
- Veldkamp, B., Verschoor, A., & Eggen, T. (2010). A multiple objective test assembly approach for exposure control problems in computerized adaptive testing. *Psicologica*, 31, 335–355.
- Wainer, H., & Mislevy, R. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computer Adaptive Testing: A Primer*: Hillsdale, NJ: Lawrence Erlbaum.
- Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450. <https://doi.org/10.1007/BF02294627>.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

